Atelier Qualité des Données du Web (QLOD'16)

Organisateurs : Samira Si-said Cherfi, Fayçal Hamdi (CEDRIC - Cnam Paris)

PRÉFACE

Dans le contexte actuel de forte compétitivité, les organisations de tous horizons sont face à un nouveau défi qui est celui de la valorisation des données. Ces données qui, grâce au développement des technologies du web, deviennent un atout stratégique. La disponibilité des données dans un format numérique et structuré a accéléré le développement des techniques d'exploration et de raisonnement sur les données dans le but d'encourager l'innovation et d'extraire de la valeur ajoutée. Ces approches ont créé un besoin pour des données réelles, fiables et de haute qualité. En conséquence, la qualité des données est devenue un enjeu important et un défi pour les années à venir.
La recherche dans ce domaine comprend des aspects théoriques, liés à la formalisation, la définition de la qualité et le développement de langages et de modèles supportant les concepts sous-jacents. Elle couvre des recherches pratiques et expérimentales par exemple le développement de méthodes d'évaluation, la proposition d'approches de validation et de benchmarks, l'expérimentation sur des jeux réels, etc.

QLOD vise à offrir un espace pour des échanges fructueux et enrichissant impliquant des chercheurs, mais aussi des professionnels du monde de l'entreprise avec une diversité de point de vue et de profil concernant la qualité des données surtout lorsqu'elles sont hétérogènes, non/peu structurées, massives et de qualités diverses.

Cette première édition présente trois travaux sur le lien entre les données et les connaissances qu'elles véhiculent, sur la qualité des données géolocalisées et sur la qualité des données liées. Ces thèmes sont complétés par une présentation de Fabian M. Suchanek autour de la construction des ontologies et de l'intégration des sources de données avec les problèmes et les opportunités de recherche sous-jacents.

Samira Si-said Cherfi
CEDRIC - Cnam Paris

Fayçal Hamdi
CEDRIC - Cnam Paris

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Nathalie Abadie
Jacky Akoka
Saïd Assar
Isabelle Comyn-Wattiau
Jérôme David
Virginie Goasdoue-Thion
Zoubida Kedad
Benjamin Nguyen
Verónika Peralta
Jolita Ralyté
Chantal Reynaud
Fatiha Saïs
Grégory Smits

TABLE DES MATIÈRES

Mise en correspondance de données textuelles hétérogènes à partir d'informations sémantiques

Nourelhouda YAHI*, Hacene BELHADEF*, Mathieu ROCHE**

* Constantine 2 Abdelhamid Mehri University, Nouvelle Ville Ali Mendjeli, Constantine, Algeria
** UMR TETIS (Cirad, Irstea, AgroParisTech) & LIRMM (CNRS, Univ. Montpellier), France

Résumé : Dans cet article, nous présentons une approche pour mesurer la similarité sémantique entre des textes hétérogènes et de qualité différente provenant de différentes sources Web. Notre approche commence par extraire le contenu des textes par deux méthodes : (i) utilisation d'un système d'extraction que nous avons implanté et qui identifie tous les mots contenus dans un texte donné, (ii) utilisation d'un thésaurus multilingue (AGROVOC). Ensuite, nous combinons les résultats des deux approches afin de mesurer la similarité entre les représentations textuelles des documents. Afin d'évaluer les résultats, nous nous appuyons sur deux ensembles de données hétérogènes issus du Web (tweets et articles scientifiques).

1. Introduction

Pour traiter les masses de données issues du Web disponibles, la problématique de recherche du *Big Data* est classiquement mise en avant avec les 3V qui la caractérisent : volume, variété et vélocité. Même si une distinction est établie entre la *véracité* (qualité) et la *variété* (hétérogénéité) des données, l'imbrication de ces deux concepts doit être prise en compte. En effet, pour avoir une connaissance exhaustive d'un sujet donné, il est nécessaire de traiter et de mettre en relation les données hétérogènes et de qualité différente. Ceci améliore indéniablement ce que nous pouvons appeler *la qualité des connaissances* traitée en prenant en compte différents points de vue. En effet, dans le domaine des SHS, une problématique tout à fait ouverte consiste à analyser des situations en considérant les multiples points de vue à travers les dires d'acteurs et d'experts. Par exemple, lorsque l'on parle de changement climatique (jeux de données traité dans cet article) plusieurs positions peuvent être mises en relief sur des supports différents (articles de presse, tweets, articles scientifiques, etc.). La qualité des connaissances est donc fortement liée à la diversité des points de vue abordés sur un sujet donné. Le lien entre *qualité des données* et *hétérogénéité* de ces dernières est donc important à considérer. Dans ce contexte, nous nous intéressons à la manière de mettre en correspondance des données textuelles hétérogènes qui sont, par nature, de qualité diverse (formats, contenus, styles linguistiques, etc.).

Plusieurs projets de recherche s'intéressent à la similarité sémantique entre des extraits de textes, mais la plupart d'entre eux s'appuient sur des textes ayant un même « niveau » linguistique et stylistique (Maguitman *et al.,* 2005) (Elsayed et Oard, 2008). Le développement d'une approche efficace qui permet de proposer une similarité sémantique entre les textes hétérogènes représente alors une problématique eminament difficile. Il existe un certain nombre de travaux de la littérature liés à l'estimation de similarité sémantique, dont beaucoup sont fondés sur l'utilisation de thésaurus. Par exemple, (Buscaldi *et al*., 2012) proposent un processus de comparaison de n-grammes sur la base d'une mesure de similarité conceptuelle utilisant WordNet[1]. Ils ont aussi appliqué une démarche similaire pour calculer la similarité sémantique de fragments textuels.

Les textes peuvent être écrits selon des styles très différents, par exemple, les tweets sont beaucoup plus difficiles à analyser linguistiquement. *A contrario,* les articles scientifiques ont une écriture plus standardisée permettant l'extraction d'information de manière plus aisée. Mais ce type de textes possède un vocabulaire de spécialité souvent plus complexe (Batista-Navarro *et al*. 2015). Dans cet article, nous proposons l'approche MIGHT (A Text Mining Process for Mapping Heterogeneous Documents) pour mesurer la similarité sémantique entre les textes hétérogènes et de qualité différente. Notre approche utilise un système extraction que nous avons implanté et qui s'appuie sur le thésaurus multilingue AGROVOC. Nous avons alors combiné les informations extraites pour calculer la similarité entre les représentations textuelles. L'article présente notre approche (section 2) qui est évaluée sur un jeu de données réel sur la thématique du changement climatique (section 3).

2. Approche MIGHT

Dans cette section, nous décrivons les détails de l'approche proposée consistant à mesurer la similarité sémantique entre deux textes de qualité différente.

Avant de mesurer la similarité entre les textes, une pondération des descripteurs linguistiques est classiquement mise en place. La pondération des termes permet d'identifier leur importance dans le texte. En général, l'idée de base est d'attribuer des poids aux termes en utilisant des informations statistiques telle que la fréquence dans un texte ou relativement à un corpus dans son ensemble (TF-IDF, Okapi, etc.).

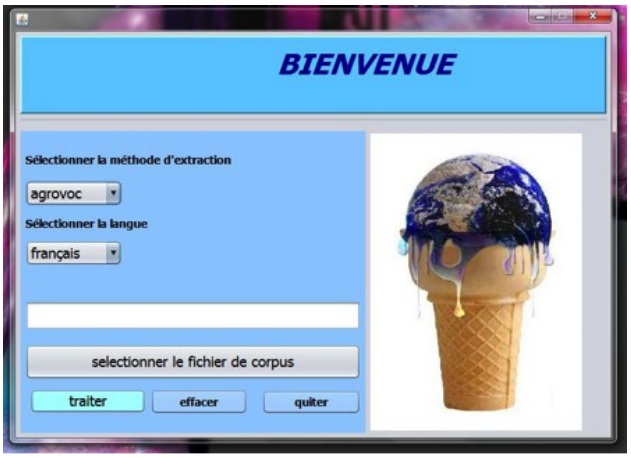
Dans notre approche, étant donné que nous traitons des textes très différents (des textes longs mais également très courts), nous avons adopté une pondération différente : nous cherchons à donner un poids plus élevé à des termes véhiculant une certaine sémantique

[1] https://wordnet.princeton.edu/

(illustré par leur appartenance à la ressource AGROVOC). Les thésaurus sont largement utilisés pour l'estimation de similarité comme WordNet qui modélise la connaissance lexicale en anglais. Dans notre approche, nous utilisons AGROVOC[2], thésaurus multilingue du domaine agronomique, qui couvre tous les domaines d'intérêt de la FAO, Organisation des Nations Unies pour l'alimentation et l'agriculture. Il est publié par la FAO et édité par une communauté d'experts. AGROVOC se compose de plus de 32.000 concepts disponibles en 23 langues. À ce jour, AGROVOC est utilisé par les chercheurs, les bibliothécaires et les gestionnaires de l'information pour l'indexation, l'extraction et l'organisation des données dans les systèmes d'information agricoles (Roche *et al*. 2015).

Les pondérations des descripteurs linguistiques ont été réalisées de la façon suivante : tous les mots identifiés dans un texte par rapport à une base représentant l'ensemble des descripteurs des corpus sont pondérés à 1, les termes qui sont extraits avec l'extracteur d'AGROVOC sont pondérés à 2 et les descripteurs linguistiques identifiés sur la base de ces deux méthodes sont pondérés à 3.

Afin d'évaluer l'approche proposée, nous avons développé un logiciel dédié (cf. figure ci-dessous).



La première étape du processus supprime tous les signes de ponctuation issus des données textuelles, puis élimine tous les « stop-words » afin d'extraire les mots, *a priori*, porteurs d'information sémantique. La deuxième étape est fondée sur l'extraction de termes

[2] http://aims.fao.org/fr/agrovoc

(mots et syntagmes) avec AgroTagger[3] ; pour cette tâche, nous nous sommes appuyés sur une classe spécifique (Maui). Le résultat de ces deux tâches est stocké dans une base de données et des vecteurs relatifs à chaque document sont construits. Enfin, une mesure de similarité (cosinus) calcule la proximité entre deux textes donnés (vecteurs pondérés).

3. Expérimentations

Dans cette section, nous montrons comment appliquer notre approche sur des ensembles de données hétérogènes (tweets et articles scientifiques) en langue française relativement à la thématique « changement climatique ».

Corpus et protocole expérimental

Dans ces expérimentations, nous avons d'abord recueilli des tweets en suivant sur les comptes Twitter des hashtags spécifiques : *#réchauffementClimatique, #changementClimatique*. Ainsi, nous avons constitué un corpus de tweets français issus d'associations, d'organisations, de célébrités et de citoyens abordant cette thématique. Puis trois autres corpus ont été utilisés. Le premier, Politweets (Longhi *et al*., 2014), rassemble des tweets de 7 personnalités issus de 6 différents groupes politiques français (34273 messages). Le deuxième corpus est une collection de résumés en français d'articles, de livres, de chapitres de livres, de thèses, etc., à partir Agritrop[4], archive ouverte du Cirad (Centre de coopération internationale en recherche agronomique pour le développement). Ces résumés scientifiques traitent du sujet du changement climatique. Le dernier corpus est une collection de résumés en français d'articles issus du laboratoire TETIS (Territoires, Environnement, Télédétection et Information Spatiale).

Ces 4 collections de données textuelles disponibles sur le Web sont notées de la manière suivante :

- CT : Tweets traitant du changement climatique.
- NCT : Tweets non liés au changement climatique (*Politweets*).
- CA : Articles scientifiques traitant du changement climatique (*Cirad*).
- NCA : Articles scientifiques non liés au changement climatique (*TETIS*).

[3] http://aims.fao.org/vest-registry/tools/agrotagger-1
[4] https://agritrop.cirad.fr/

Résultats

Les expérimentations sont réalisées selon 10 itérations, pour chacune d'elle, nous avons sélectionné aléatoirement N éléments (N = 10) – le nombre d'exécutions total est de 3000. Ensuite, nous appliquons la similarité entre les éléments (les documents).

Le tableau ci-dessous montre les résultats obtenus (moyenne des similarités pour chaque itération) afin de comparer (i) CT et CA, (ii) CT et NCT, (iii) CT et NCA. Les degrés de similarité les plus élevés pour cinq itérations est la similitude entre CT et CA (couvrant des sujets proches) mettent en avant des premiers résultats encourageants restitués par notre approche MIGHT. Dans la suite des travaux, il sera nécessaire d'analyser d'un point de vue qualitatif les résultats obtenus, en particulier les faux positifs et faux négatifs obtenus. Ceci permettra de discuter dans quelle mesure les mots originaux issus des textes ou les termes d'Agrovoc ont permis d'améliorer ou non les différentes mises en correspondance des documents.

Iteration	CT/NCT	CT/CA	CT/NCA
1	0.0073	0.0025	0.0078
2	0	0.0063	0
3	0	0.0128	0.0093
4	0	0	0
5	0	0	0
6	0	0.0182	0.032
7	0	0.0069	0
8	0	0.0182	0.032
9	0	0.0069	0
10	0	0.0106	0

4. Conclusion

Dans cet article, nous avons abordé la question de l'évaluation de la similarité sémantique entre des documents de nature différente mais qui peuvent porter sur des sujets proches. Notre approche est fondée sur l'extraction de descripteurs linguistiques issus d'un texte (mots) et des termes (mots et syntagmes) propres à un thésaurus en appliquant une pondération « sémantique » spécifique. Notre méthode a tendance à rapprocher des textes ayant des thématiques proches ce qui permet de mettre en relation des données de qualité différente. De nombreuses perspectives peuvent être proposées comme (i) l'élimination du vocabulaire spécifique aux tweets (phrases ou expressions spécifiques), (ii) l'expansion de contextes (par exemple, en considérant l'ensemble de tweets écrits par le même auteur dans une même fenêtre temporelle).

Remerciements

Ce travail est financé par la Région Languedoc-Roussillon et par les Fonds Européens de Développement Régional (**projet SONGES** : http://textmining.biz/Projects/Songes).

Références

R. T. Batista-Navarro, R. Rak, S. Ananiadou (2015). Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminformatics 7(S-1): S6*

D. Buscaldi, R. Tournier, N. Aussenac-Gilles, J. Mothe (2012). Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In F*irst Joint Conference on Lexical and Computational Semantics (SEM)*, pages 552-556, Montreal, Canada.

T. Elsayed, J. Lin, D. W. Oard (2008). Pairwise document similarity in large collections with mapreduce. In P*roceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*: Short Papers, Columbus, Ohio, 2008, pp.2652-68.

J. Longhi, C. Marinica, B. A. Alkhouli (2014). Polititweets : corpus de tweets provenant de comptes politiques influent. In *Chanier T. Banque de corpus CoMeRe*. Ortolang.fr.

A.G. Maguitman, F. Menczer, H. Roinestad, A. Vespignani (2005). Algorithmic detection of semantic similarity. In *Proceedings of 22nd International Conference on World Wide Web*, Chiba, Japan, 2005, pp. 107-116.

M . Roche, S. Fortuno, J.A. Lossio-Ventura, A. Akli, S. Belkebir, T. Lounis, S. Toure (2015). Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cahiers Agricultures*. Volume 24, numéro 5, p.313-320